

## 1 Introduction

La langue arabe s'écrit et se lit de droite à gauche, et fait partie du groupe des langues sémitiques. Ce groupe se divise en langues sémitiques orientales, sémitiques occidentales et sémitiques méridionales. [25] L'arabe existe et se développe à partir du 7ème siècle grâce à la propagation de l'islam et la diffusion du Coran qui est considéré comme la base de cette langue.

Les recherches sur le traitement automatique de l'arabe ont débuté vers les années 1970, et les premiers travaux concernaient notamment les lexiques et la morphologie. Ainsi, la croissance rapide du World Wide Web (WWW) et l'explosion primordiale des ressources de l'indexation et des outils de recherche, influent effectivement sur le développement dans le domaine de la recherche d'information. [25]. La langue arabe a une morphologie complexe et par ses propriétés elle est considérée comme une langue difficile à maîtriser dans le domaine de RI.

Dans ce chapitre nous présenterons la morphologie de la langue arabe ainsi que certains problèmes liés à ses propriétés.

## 2 Morphologie de l'arabe

### 2.1. Les lettres

La langue arabe est composée de 28 lettres (table 2.1).

Lettre arabe	Correspondant français	Prononciation	Lettre arabe	Correspondant français	Prononciation
ا	a	Alef	ض	d	Dad
ب	b	Ba'	ط	t	Tah
ت	t	Ta'	ظ	z	Zah
ث	th	Tha'	ع	'	Ayn
ج	j	Jim	غ	gh	Ghayn
ح	h	Hha'	ف	f	Fa
خ	kh	Kha'	ق	q	Qaf
د	d	Dal	ك	k	Kaf
ذ	d	Thal	ل	l	Lam
ر	r	Ra	م	m	Mim
ز	z	Zayn	ن	n	Nun
س	s	Sin	ه	h	Ha
ش	sh	Shin	و	w	Waw
ص	s	Sad	ي	y	Ya

**Table 2.1 :** Les Lettres Arabes. [9].

Les lettres changent leur forme de présentation selon leur position (au début, au milieu ou à la fin du mot) et ils sont connectés sauf (ا, و, ز, د, ذ) qui ne se joint pas à gauche.

Le Tableau suivante montre les changements de la lettre (غ) selon leur position dans le mot.

à la fin d'une lettre non joignable	à la fin	au milieu	au début
فراغ	بليغ	الغابة	غيم

**Table 2.2 :** Les Changements de la Lettre غ.

## 2.2 Les voyelles

Un mot arabe s'écrit avec des consonnes et des voyelles. Les voyelles sont ajoutées au-dessus ou au-dessous des lettres. Elles sont nécessaires à la lecture et à la compréhension correcte d'un texte, elles permettent de différencier des mots ayant la même représentation. [9]

Il y a trois voyelles brèves et sept signes orthographiques qui s'ajoutent aux consonnes. Ces trois voyelles brèves sont :

- Fatha « َ » : elle surmonte la consonne et se prononce comme un «a» français ;
- Damma « ُ » : elle surmonte la consonne et se prononce comme un «ou» français ;
- Kasra « ِ » : elle se note au-dessous de la consonne et se prononce comme un « i » français).

Les sept signes orthographiques sont :

- Sukun « ْ » : ce signe indique qu'une consonne n'est pas suivie (ou muet) par une voyelle. Il est noté toujours au-dessus de la consonne.

→ Les trois signes de tanwin : lorsque les trois voyelles brèves sont doublées, elles prennent un son nasal, comme si elles étaient suivies de «n» et on les prononce respectivement :

- an « ً » pour les Fathatan.
- in « ٍ » pour les Kasratan.
- un « ٌ » pour les Dammatan.

→ Chadda « ّ » renforcer une consonne quelconque.

→ Wasla « ِْ » quand la voyelle Alif au commencement d'un mot doit être absorbée par la dernière voyelle du mot qui précède.

→ Madda « ّ » la madda (prolongation) se place sur Alif pour indiquer que cette lettre tient lieu de deux alifs consécutifs ou qu'elle ne doit pas porter le Hamza.

## 2.3 Structure d'un mot

En arabe un mot peut représenter toute une proposition. La représentation suivante schématise une structure possible de mot complexe : [25]

Antéfixe	Préfixe	Noyau	Suffixe	Post fixe
----------	---------	-------	---------	-----------

**Table 2.3 :** Structure d'un Mot Arabe.

### 2.3.1 Les antéfixes [14]

Les antéfixes sont généralement des prépositions agglutinées au début des mots. Ils se combinent entre eux pour donner les traits syntaxiques, coordonnant, terminant ...etc.

### 2.3.2 Les préfixes [14]

Les préfixes, habituellement représentés par une seule lettre, indiquent la personne de conjugaison des verbes au présent. Ils ne se combinent pas entre eux.

### 2.3.3 Les suffixes [14]

Les suffixes sont les terminaisons de conjugaison des verbes et de marques duelles/plurielles/femelles pour les noms y compris les adverbaux. Ils ne se combinent pas entre eux.

### 2.3.4 Les post fixes [14]

Finalement, les post fixes représentent des pronoms attachés à la fin des mots.

Exemple : "أنتفكروننا؟" ce mot exprime la phrase en français: «*Est ce que vous vous souvenez de nous ?*». La segmentation de ce mot donne les constituants suivants (Figure 2.1) :

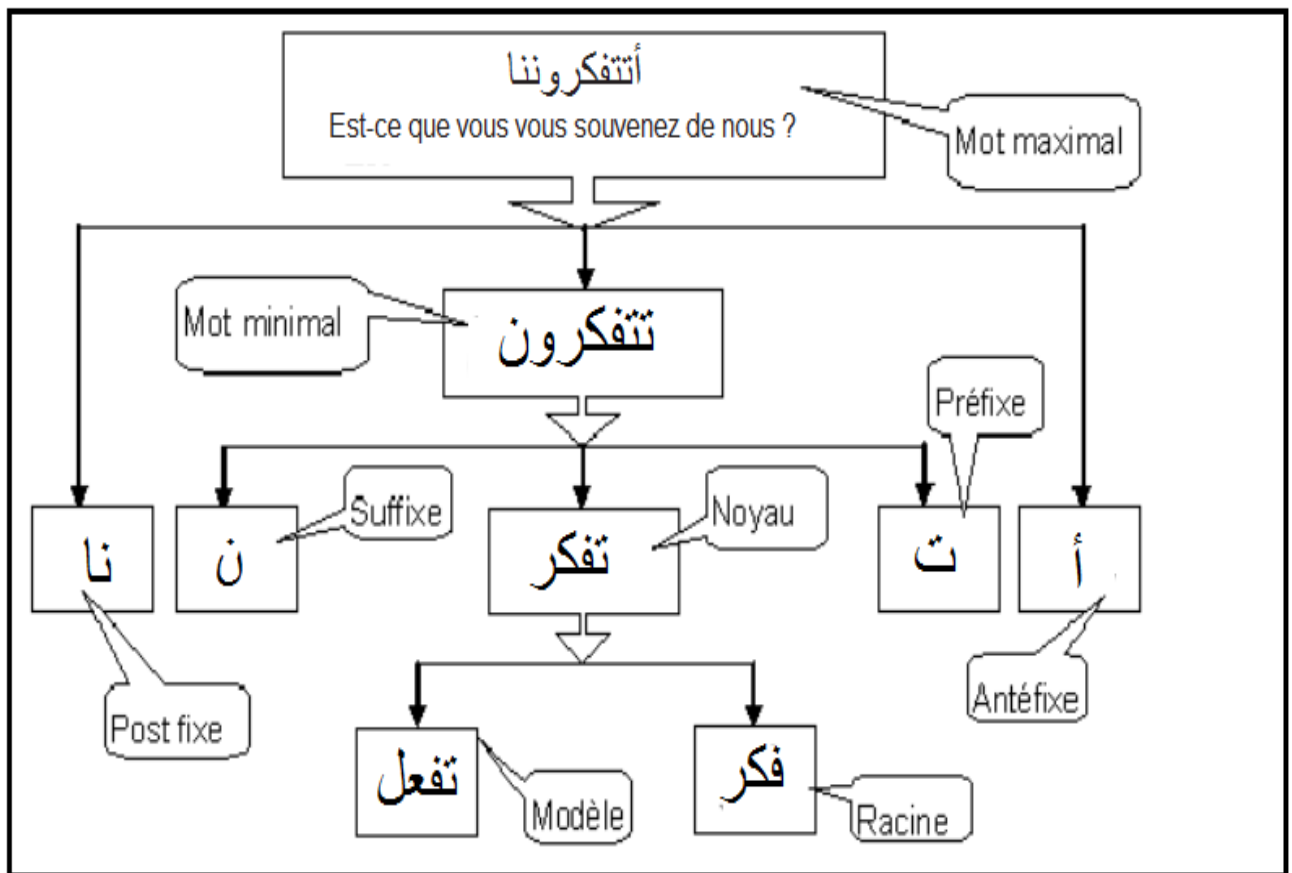


Figure : 2.1 : Segmentation d'un mot en arabe.

## 2.4 Catégories d'un mot

La langue arabe comporte trois catégories de mots :

### 2.4.1 Les verbes

En arabe, la majorité des mots dérivent d'un verbe de 3 consonnes qui représente une racine d'un groupe de mots. Comme en français, le mot en arabe se détermine à partir d'un radical en rajoutant des préfixes, des suffixes ou les deux en même temps. [7]

Comme les autres langues, la conjugaison de verbe en arabe dépend de facteurs suivants :

- L'aspect : accompli (passé) ou inaccompli (présent).
- Le nombre du sujet : singulier, pluriel ou duel.
- Le genre du sujet : masculin ou féminin.
- La personne : première, deuxième ou troisième.
- La voix : active ou passive.

Exemple : La combinaison de ces trois consonnes (أ+ك+ل) produit le verbe أكل / A-k-l (manger).

La conjugaison des verbes se fait en rajoutant des préfixes, des suffixes ou les deux. La langue arabe possède trois temps :

- L'accompli : indique le passé et les verbes conjugués se distinguent par des suffixes. Pour notre exemple, avec le féminin pluriel, on obtient / أكلن / « elles ont mangé » ; pour le masculin pluriel, on obtient / أكلوا / « ils ont mangé ».
- L'inaccompli présent : les verbes conjugués à ce temps se distinguent par les préfixes. Pour notre exemple, au masculin singulier on obtient / يأكل / « il mange » ; et pour le féminin singulier on obtient / تأكل / « elle mange ».
- L'inaccompli futur : la conjugaison d'un verbe au futur nécessite d'ajouter l'antéposition au début du verbe conjugué à l'inaccompli. En ajoutant l'antéposition à notre exemple س on obtient سيأكل / « il mangera », qui désigne le futur ; on peut également ajouter l'antéposition سوف on obtient سوف يأكل / « il va manger ».

#### 2.4.2 Les noms

En arabe les noms sont divisés en deux familles, ceux qui sont dérivés à partir d'une racine (verbale) et les autres comme les noms étrangers et certains noms fixes. La première famille est composée des tous les noms qui sont dérivés à partir d'une racine verbale. [7] La variabilité des noms obéit à plusieurs règles, en ajoutant des morphèmes spécifiques :

- ✓ Le féminin singulier : pour obtenir le nom féminin singulier, dans la majorité des cas on ajoute le lettre ة (exemple : « قط /chat» devient « قطة /chat»).
- ✓ Le féminin pluriel externe : pour obtenir le nom féminin pluriel on ajoute les deux lettres ات exemple : « رسام devient رسامات ».
- ✓ Le masculin pluriel externe : on ajoute les deux lettres ون ou ين qui dépendent de la

position du nom dans la phrase (avant ou après le verbe). Exemple : « معلمي/enseignant » se transforme en « معلمون ou معلمين /enseignants ».

✓ Le pluriel masculin, féminin et interne : c'est le cas le plus complexe en arabe, la construction de ces types des noms s'obtient en insérant des lettres au début, au milieu ou à la fin (exemple : « عمل/travail » se transforme en « أعمال/travaux » et « قلب/cœur » se transforme en « قلوب/ cœurs »).

Notons que, la morphologie des noms arabes dépend de ces fonctions.

### 2.4.3 Les particules

Les particules sont les mots outils pour une langue donnée. En arabe, comme le notent : [13] « Les particules sont classées selon leur fonction dans la phrase, on en distingue plusieurs types (introduction, explication, conséquence). Elles jouent un rôle important dans l'interprétation de la phrase ».

- Préposition : exemple حَتَّى، عَنْ، فِي، ل، ك، ب.
- Particules de coordination : exemple وَ، ف، ثُمَّ، أَوْ.
- Particules interrogatives : exemple مَاذَا، هَلْ، أ.
- Particules d'affirmation : exemple نَعَمْ، بَلَى، أَجَل.
- Particules de négation : exemple لَا، لَنْ.
- Particules distinctive : exemple أَي.
- Particules relatives : exemple الَّذِي، الَّتِي،.
- Particules de futur : exemple سَوْفَ، س.
- Particules conditionnelles : exemple إِنْ، لَوْ.

Les particules peuvent avoir des préfixes et suffixes ce qui produit un problème de complexité quant à leur identification dans l'étape de segmentation.

## 3 Complexité du traitement automatique de l'arabe

### 3.1 L'ambiguïté

L'ambiguïté des langues est l'une des questions la plus complexe dans le domaine de RI. Notons que Le taux de l'ambiguïté dans la recherche d'information en langue arabe est considéré parmi les plus élevés des langues pratiquées [14]. Il existe différents types d'ambiguïtés. On citer :

- Les variations orthographiques sont très répandues en arabe, certaines combinaisons des lettres peuvent être écrites de différentes manières. La présence de l'une ou de l'autre de ces accentuations peut introduire une sorte d'ambiguïté. [14].  
→ Exemple la lettre « أ » peut être écrite de différentes manières أ, آ, إ
- Un mot sans voyelles peut générer plusieurs cas d'ambiguïtés lexicales.  
→ Variations lexicales comme dans le cas de « الراحة : كف اليد » et « الاسترخاء : الراحة »
- La phrase arabe est caractérisée par une grande variabilité au niveau d'ordre de ses mots. Cette variabilité de l'ordre des mots, provoque des ambiguïtés syntaxiques.  
→ Exemple : « سأل الطالب الأستاذ/سأل الأستاذ الطالب ».

### 3.2 Problème de la forme et capitalisation

En plus du phénomène d'ambiguïté, il y a un autre problème de la forme.

- L'arabe a une morphologie très compliquée. un mot arabe peut avoir une forme plus compliquée s'il y a présence de tous ces affixes attachés à sa forme standard. [14].  
→ Exemple : « سألتُمُونِهَا »
- Les pluriels cassés sont communs. Les pluriels cassés souvent ne ressemblent pas à la forme singulière, ils n'obéissent pas à des règles morphologiques normales. [25].  
→ Exemple : « غراب - غرابان ».
- De plus la capitalisation n'existe pas dans l'arabe ce qui rend la reconnaissance des noms propres, des acronymes, et des abréviations très difficile [25].  
→ Exemple : « nom propre : احمد » et « le verbe je remercie : احمد »

## 4 Conclusion

La langue arabe a une morphologie riche, complexe et difficile à traiter, et à cause de leur complexité elle produit un haut niveau d'ambiguïté qui constitue en fait le problème majeur du traitement de la langue arabe dans le domaine de la recherche d'information. Dans ce chapitre, nous avons présenté certaines caractéristiques de la langue arabe.